What Works Clearinghouse Intervention Rating Scheme

Factors Determining the Rating

Explicit heuristics will be applied to support two judgments about the findings of each qualifying study about a given outcome (or outcome domain) for a given intervention:

1. The direction, magnitude, and statistical significance of the empirical effect estimate. This will be characterized as a ***statistically significant positive, substantively important positive***, ***indeterminate***, or ***statistically significant negative*** effect.
2. The quality of the research design generating the effect estimate. This will be characterized as a ***strong*** or ***weak*** design. (See the WWC Study Review Standards for further details.)

The rating scheme based on these two factors is presented below. After that are the detailed descriptions and heuristics for making the judgments on these factors for each study and outcome.

Rating Scheme Based on These Judgments

Positive Effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Two or more studies showing *statistically significant positive* effects, at least one of which met WWC evidence standards for a *strong* design.
- No studies showing *statistically significant* or substantively important *negative* effects.

***Potentially Positive Effects*:** Evidence of a positive effect with no overriding contrary evidence.

- At least one study showing a *statistically significant* or *substantively important positive* effect.
- No studies showing a *statistically significant* or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing *statistically significant* or *substantively important positive* effects.

Mixed Effects: Evidence of inconsistent effects as demonstrated through either of the following.

- At least one study showing a *statistically significant* or *substantively important positive* effect; AND at least one study showing a *statistically significant* or substantively important *negative* effect, but no more such studies than the number showing a *statistically significant* or *substantively important positive* effect.
- OR, at least one study showing a *statistically significant or substantively important* effect AND more studies showing an *indeterminate* effect than showing a *statistically significant or substantively important* effect.

No Discernible Effects: No affirmative evidence of effects.

- None of the studies shows a *statistically significant or substantively important effect, either positive or negative.*

Potentially Negative Effects: Evidence of a negative effect with no overriding contrary evidence.

- At least one study showing a *statistically significant* or substantively important *negative* effect.

- No studies showing a *statistically significant* or *substantively important positive* effect OR more studies showing *statistically significant* or substantively important *negative* effects than showing statistically *significant* or *substantively important positive* effects.

Negative Effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Two or more studies showing *statistically significant negative* effects, at least one of which is based on a *strong* design.
- No studies showing *statistically significant* or *substantively important positive* effects.

Evidence Base and Heuristic Rules

Points of Evidence

For each study of the intervention and for each outcome, the following points of evidence are the basis for characterizing the empirical findings:

1. Quality of the study design: RCT (meets evidence standards) or QED (meets evidence standards with reservations) under the current WWC criteria.
2. Effect size: A single effect size or, in the case of multiple measures of the specified outcome, either (i) the mean effect size, or (ii) the effect size for each individual measure within the domain. The effect size is defined as the standardized mean difference (i.e., the difference between the student-level posttest means on an outcome variable divided by the pooled standard deviations, either calculated directly or derived from other appropriate statistics, corrected for small sample sizes).
3. Sample size: The number of units of assignment per condition and the number of students in those units per condition if students were not the units of assignment.
4. Statistical significance of the effect based on a correct ("aligned") analysis if reported. Statistical significance is assumed to mean the conventional alpha=.05, two-tailed for single measures and for mean effects within each domain. When multiple hypothesis tests are performed using the number of measures greater than one (m>1measures) within each domain, the Benjamini Hochberg procedure may be used to correct for multiple comparisons and identify statistically significant effects for individual measures (Benjamini, Y., and Y. Hochberg, *Journal of the Royal Statistical Society* 1995, Vol. 57, No.1, 289-300 [http://www.math.tau.ac.il/~ybenja/MyPapers.html]).

Characterizing the Quality of the Research Design Generating the Effect Estimates

The heuristics for categorizing the quality of the research design used in a given study are as follows:

- ***Strong design:*** designs that meet the WWC's evidence standards, which are RCTs without severe design or implementation problems.)
- ***Weak design:*** designs that meet WWC's evidence standards with reservations, which include RCTs with severe design or implementation problems, and QEDs with equating and without severe design or implementation problems.

(See the WWC Study Review Standards for further details.)

Characterizing the Direction and Magnitude of the Empirical Effect Estimate

These heuristics are applied to the outcome variable(s) identified by the Principal Investigator (PI) as relevant to the review. The PI may choose to ignore some variables if they are judged

sufficiently peripheral or unrepresentative and consider only the remaining ones. Similarly, if the PI judges that there is one core variable with all the others secondary or subsidiary, only that one may be considered.

A. Definitions and Suggested Defaults

The heuristics in the next section require that values be set for certain terms. These terms and associated procedures are defined below with suggested default values.

Minimum effect size. The smallest positive value at or above which the effect is deemed substantively important with relatively high confidence for the outcome domain at issue. Effect sizes at least this large will be taken as a qualified positive effect even though they may not reach statistical significance in a given study. The suggested default value is a student-level effect size greater than or equal to 0.25 ($ES \geq 0.25$), corresponding to a 10 percentile point difference between the percentile rank of the average student in the comparison group (50^{th} percentile) and the percentile rank of the average student in the intervention group (60^{th} percentile) based on the comparison group distribution. The PI may set a different default if explicitly justified in terms of the nature of the intervention or the outcome domain. A similar default applies in the negative direction. The suggested default value for a minimum negative effect is a student-level effect size less than or equal to –0.25 ($ES \leq -0.25$).

t test adjusted for clustering. A *t* test applied to the effect size (or mean effect size in cases of multiple measures of the outcome) that incorporates an adjustment for clustering. This procedure allows the reviewer to test the effect size directly in cases where a misaligned analysis is reported. (Computational details are provided in the appendix.) However, the clustering adjustment requires specifying an ICC value. The suggested default ICC value for achievement outcomes is .20. The suggested default ICC for behavioral and attitudinal outcomes is .10. The PI may set different defaults if explicitly justified in terms of the nature of the research circumstances or the outcome domain.

B. Heuristics for Characterizing Effects of a Study

(Note: The italicized terms involve default values and are defined above.)

Statistically significant positive effect: Any one of the following:

If the analysis as reported by the study author is properly aligned:

- For a single outcome measure within an outcome domain, either of the following is appropriate. (If the results differ, select the strategy which demonstrates significance.)
  - The effect is reported as positive and statistically significant.
  - The effect size is positive and statistically significant when tested using a *t test adjusted for clustering.*
- For multiple measures of outcomes within an outcome domain, any of the following are appropriate. (If the results differ, select the strategy that demonstrates statistical significance.)
  - Univariate statistical tests are reported for each outcome measure and *at least half* of the effect sizes are positive and statistically significant and *no* effect sizes are negative and statistically significant, ignoring multiple hypothesis tests.

- The omnibus effect for all the outcome measures together is reported as positive and statistically significant on the basis of a multivariate statistical test.
- Univariate statistical tests are reported for each outcome measure and the effect size for *at least one* measure within the domain is positive and statistically significant and *no* effect sizes are negative and statistically significant, *when accounting for clustering and for multiple hypothesis tests within the domain*.
- The *mean* effect size for the multiple measures of the outcome is positive and statistically significant when tested using a *t test adjusted for clustering.*[1]

If the analysis as reported by the study author is not properly aligned, either of the following is appropriate:

- The effect size or the *mean* effect size (if multiple measures of outcomes within a domain) is positive and statistically significant when tested using a *t test adjusted for clustering*.
- Univariate statistical tests are reported for each outcome measure and *at least one* effect size is positive and statistically significant and *no effect sizes* are negative and statistically significant, *accounting for clustering and multiple comparisons within the domain*.

Substantively important positive effect:

- The effect size is not statistically significant in any of the senses described above, but the student-level effect size (if there was a single student-level measure within an outcome domain) or the *mean* effect size based on multiple student-level findings (if there were multiple student-level measures within an outcome domain) is equal to or greater than the *minimum effect size*.[2]

Indeterminate effect:

- The effect size is not statistically significant and does not qualify as a substantively important positive effect as defined above (that is, the effect size or the mean effect size is less than the *minimum effect size*).

Substantively important negative effect:

- The effect size is not statistically significant in any of the senses described above, but the student-level effect size (if there was a single student-level measure within an outcome domain) or the *mean* effect size based on multiple student-level findings (if there were multiple student-level measures within an outcome domain) is equal to or less than the *minimum negative effect size*.[2]

[1] Note that this formula is still acceptable if there is no clustering, as the clustering term drops out of the equation.
[2] Note that this criterion, as well as the default *minimum effect size*, is entirely based on student-level ESs. Cluster-level ESs are ignored for the purpose of the rating scheme because they are based on a different ES metric than the student-level ESs, and therefore not comparable with student-level ESs. Moreover, cluster-level ESs are relatively rare, and there is not enough knowledge in the field yet to set a defensible *minimum effect size* for cluster-level ESs.

Statistically significant negative effect: Any one of the following where no statistically significant or substantively important positive effect has been detected (in the sense outlined above):

If the analysis as reported by the study author is properly aligned:

- For a single outcome measure within an outcome domain, either of the following is appropriate. (If the results differ, select the strategy that demonstrates significance.)
  - The effect is reported as negative and statistically significant.
  - The effect size is negative and statistically significant when tested using a *t test adjusted for clustering.*
- For multiple measures of outcomes within an outcome domain, any of the following is appropriate. (If the results differ, select the strategy which demonstrates significance.)
  - Univariate statistical tests are reported for each outcome measure and *at least half* of the effect sizes are negative and statistically significant and *no* effect sizes are positive and statistically significant, ignoring multiple hypothesis tests.
  - The omnibus effect for all the outcome measures together is reported as negative and statistically significant on the basis of a multivariate statistical test.
  - Univariate statistical tests are reported for each outcome measure, and *at least one* effect size is negative and statistically significant and *no* effect sizes are positive and statistically significant, *accounting for clustering and multiple comparisons within the domain*.
  - The *mean* effect size for the multiple measures of the outcome is negative and statistically significant when tested using a *t test adjusted for any clustering.*[3]

If the analysis as reported by the study author is not properly aligned, either of the following is appropriate:

- The effect size or the *mean* effect size (if multiple measures of outcomes within an outcome domain) is negative and statistically significant when tested using a *t test adjusted for clustering*.
- Univariate statistical tests are reported for each outcome measure and *at least one effect size* is negative and statistically significant and *no* effect sizes are positive and statistically significant, *accounting for clustering and multiple comparisons within the domain*.

[3] Note that this formula is still acceptable if there is no clustering, as the clustering term drops out of the equation.

Appendix: Computational details for the *t test adjusted for clustering*

To determine if it is plausible that the effect size in a study with a misaligned analysis is statistically significant

(1) The reviewer has:

N_T, N_C, and $N = N_T + N_C$
(student-level sample sizes for the intervention and comparison groups respectively)

$m = m_T + m_C$
(number of clusters—classrooms or schools)

$ES = (X_T - X_C)/S_p$
(effect size computed from student level means and SDs with no attention to clustering)

(2) A default rho is assumed (current default is ρ=.20)

(3) The *t* statistic is computed for the effect size ignoring clustering:

$$t = ES\sqrt{\frac{N_T N_C}{N_T + N_C}}$$

(4) The *t* value above is corrected for clustering using the default rho and assuming equal *n* in each cluster:

$$t_A = ct \quad \text{where } c = \sqrt{\frac{(N-2) - 2\left(\frac{N}{m} - 1\right)\rho}{(N-2)[1 + \left(\frac{N}{m} - 1\right)\rho]}}$$

(5) Adjusted degrees of freedom are calculated:

$$h = \frac{\left[(N-2) - 2\left(\frac{N}{m} - 1\right)\rho\right]^2}{(N-2)(1-\rho)^2 + \frac{N}{m}\left(N - 2\frac{N}{m}\right)\rho^2 + 2\left(N - 2\frac{N}{m}\right)\rho(1-\rho)}$$

(6) Significance is determined in the usual way using adjusted t_A with adjusted *df=h*